International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Multi-Microphone Speech Dereverberation and Noise Reduction using Long Short-Term Memory Networks

Seema Arote[1], Vijay Mane*[2], Dattaray Bormane[3] and Shakil S. Shaikh[4]

Submitted: 10/09/2023 **Revised**: 15/11/2023 **Accepted**: 25/11/2023

Abstract: In the field of speech signal analysis, deep learning has recently demonstrated substantial advantages. In contrast to the other techniques for dereverberation, traditional speech dereverberation approaches underperform when dealing with severe reverberation, especially when dealing with noise and variations in the source to array distance. This work suggested an enhanced reinforcement learning (RL) as the basis for a novel approach to speech dereverberation and denoising. In this method, speech components that are reverberant and noisy (like the logarithmic spectrum) are mapped with basic coefficients for clear speech by using a Long Short Term Memory (LSTM) that is learned using a clean/distorted speech corpus. The suggested method is intuitive and effective in reducing the distorting effects of reverberation and ambient noise. Numerous experiments demonstrate that the suggested method significantly boosts the predicted speech perception and quality in noisy environments. The source-to-array distance, reverberation time (RT), noise variety, and signal-to-noise ratio (SNR) are just some of the variables that are tested. We find that our technique significantly outperforms the competitor. It consistently outperforms all benchmark techniques and can improve speech quality even in low-SNR environments

Keywords: *Reverberation, Dereverberation, Room Impulse Response (RIR), Long Short Term Memory Network (LSTM), Reinforcement Learning (RL), Signal to noise ratio (SNR)*

1. Introduction

Human communication is most natural through speech; individuals can't interact or share information effectively without it. During the pandemic, study, work, and collaboration were primarily reliant on videoconference or web-based meeting systems, with speech representing as the primary mode of communication. Due to the challenges posed by applications like mobile speech communication devices, cochlear implants, and reliable speech recognition, commitment to lifelong learning in speech dereverberation and denoising has always been consistently high. When one or more microphones placed at a distance from the speaker capture speech signals in an enclosed space, the voice signal was delayed and dampened, and the signal that was observed is a superposition of these replicas, which were reflected off of nearby walls, roofs, and floors [1]-[2]. As a result, speech superior and clarity suffer, especially when reverberation consequences are severe [3]. Reverberation has a significant impact on the superiority and clarity of speech. The overall performance of automatic speech recognition (ASR) frameworks may also suffer immensely in reverberant conditions, particularly when the training and test phases are out of sync. When reverberation is excessive, hearing in humans is harmed from significant interference [4]. Speech enrichment is process of processing speech that has been distorted to achieve fundamentally clear speech, consequently speech quality enhancements [5]. Weighted prediction error (WPE) traditionally utilizes a linear regression algorithm to calculate late reverberation and afterwards revoke it. They use the power spectrum of early speech for this. Its estimation, however, is based on an evolutionary algorithm with a significant amount of computation. Another issue is that the WPE makes the assumption that there is no noise. As a result, performance suffers in noisy environments [6]. Supervised speech enhancement techniques have developed as a result of deep learning has advanced rapidly in recent years. Two of these methods are especially important. The mapping [7] is a technique that uses a deep neural network for predicting clear speech features directly from distorted speech features. Normally, this function applies to the logarithmic power spectrum, which also follows human auditory rules and aids in DNN learning. The other method is masking [8], which forecast distorted speech and produces clean speech from the estimated a transitional state and the predicted intermediate state based on the feature of distorted speech. The reverberation components of reverberant speech can be significantly reduced by mapping and masking techniques because of DNN's capacity for learning. Among the masking techniques are the ideal binary mask [9], ideal ratio mask [10], ideal amplitude mask [11], phase-sensitive mask (PSM) [12], and complex ratio mask [13]. The power of the distorted speech is represented by IRM using the discrepancy within clean speech and additive noise, which

[1,2] *Department of Electronics and Telecommunication Engineering Vishwakarma Institute of Technology, Pune, India*
[3] *Department of Electronics and Telecommunication Engineering AISSMS, College of Engineering, Pune, India*
[4] *Department of Electronics and Computer Engineering Pravara Rural Engineering College Loni, India*
**Corresponding Author Email: vijay.mane@vit.edu*

restricts the IRM training target to [0; 1], as a sum of both the original clean speech and additive noise. This makes DNN training easier and advances IRM's accomplishment of decreasing additive noise. With the increasing amount of noise and reverberation, framework illustrates for dereverberation and denoising will undoubtedly help listeners perceive speech more clearly in both typical listening situations and many speech enhancement software applications. The significance of the problem has led to a lot of research effort being put into improving speech that has been distorted by ambient noise or room reverberation in recent years. For reverberation, the previous frames usually influence the current speech. Most speech dereverberation methods in traditional neural networks focus on the current moment of speech and can only use short-term clues to clean up the sound or learn a filter to reduce background echoes. They struggle to effectively remove echoes that happen later in the audio, making their dereverberation performance less effective. To tackle this problem, this research suggests using a special type of neural network called LSTM, combined with reinforcement learning, to better grasp important signals in the audio. First, we investigate the efficacy using pre-emphasis Filter-based data, followed by a high-dimensional representation to focus on a more useful time index.

2. Literature Review

Z. Wang and colleagues [14] investigate how deep learning can be used for both single and multi-channel voice dereverberation. They extend the use of magnitude-domain masking and mapping-based methods to a more complex approach for single-channel processing. In this approach, deep neural networks are trained to predict the real and imaginary components of the direct sound signal from the reverberated and noisy ones, which can be quite challenging due to the complexity of the audio data. To clean up the direct sound in multi-channel processing, they first use a minimum variance distortion less response beamformer. Then, they take the real and imaginary components of the delayed signal, which they've predicted to be a filtered version of unwanted noise, and use these as additional information to improve the dereverberation process. H. Chen et al. [15] propose a special neural network with two separate information streams. This network is designed to handle tasks related to reducing echoes and separating voices in audio, especially when there are different numbers of voices in the audio. Deep attractor networks (DANs) use discriminative embeddings and speaker attractors to separate audio. H. Wang et al. [16] investigates an efficient method for leveraging environmental knowledge to enhance voice dereverberation efficacy in real-world reverberant settings. They suggested a deep neural network (DNN)-based temporal-contextual attention method for context speech dereverberation that can adjust to contextual input.

Y. Fu et al. [17] propose a multi-channel network that can clean up echo from speech, improve the audio quality, and separate different voices all at the same time. They use a technique called "attentional selection" of multi-channel features, which is inspired by methods used in various audio processing approaches like end-to-end unmixing, fixed-beamforming, and extraction methods. They present a new deep complex convolutional recurrent network as the framework of audio unmixing, and a neural network-based weighted prediction error is cascaded beforehand for speech dereverberation. C. Fan et al. [18] propose a method for cleaning up audio by removing both noise and echoes at the same time. They achieve this by using a technique called "deep clustering" and training the system with deep embedding attributes, which helps improve the audio quality. The deep clustering network used to retrieve deep embedding features that do not contain any noise or unwanted sound. during the denoising step. The anechoic voice and leftover sound signals are used to create these embedding characteristics. They can depict the wanted signals' assumed spectrum filtering patterns, which are discriminative characteristics. Instead of using the unconstrained K-means grouping method at the dereverberation step, another trained neural network to forecast the clean, echo-free speech based on these deep embedding attributes.

H. Li et al. [19] suggest a neural network for predicting the power spectral distribution of early speech, as well as a binary filter for distinguishing target speech from background noise. To reduce the effect of noise on echo path prediction, a dual-filter approach is used to simulate the echo pathways of target speech as well as background noise separately. Xiao et al. [20] proposed a technique called "linear feature adaptation." They use a special transformation called the "cross transform" to change many sets of voice data into a different feature space. When you have a model representing clear speech data, this transformation is computed to make sure that the altered speech data has the highest possible chance of being accurate. Unlike the DNN method, no concurrent data is used, and no assumptions are made about deformation kinds. K. Han et al. [21] suggest using supervised learning to conduct voice dereverberation, and the supervised method is then expanded to handle both dereverberation and denoising. Deep neural networks have been taught to learn a spectrum translation from a distorted speech magnitude spectrogram to a clear speech magnitude spectrogram. Y. Masuyama et al. [22] suggest a unique end-to-end design that combines dereverberation, beamforming, and self-supervised learning modelling inside one single neural network. Sheeja, J. et al. [6] suggested a new voice segmentation and dereverberation

technique based on the integration of Principal Component Analysis based on Locally Weighted Projection Regression and Weighted Forecast Error based on Deep Neural Network. Prior to applying Blind Source Separation and Blind Dereverberation, the suggested technique pre-processes the combined reverberant output. S. Gul et al. [23] propose a novel approach for improving the sound quality of a single voice source using two ear-like devices. These devices can detect differences in the sound between your ears, both from the direct sound and the echoes sounds. To understand the echoey sounds better, they use two special devices positioned apart from each other. They then train a deep learning network called U-Net to learn from the differences in the sounds detected by these devices, which helps enhance the quality of the audio.

3. Methodology

The proposed method introduces technique for building systems that improve the clarity and quality of speech in difficult, noisy, and echoing environments. This method allows each moment of speech to understand the whole sequence without making the process too complicated. They incorporate a design called "RL" (Reinforcement Learning) into their framework for reducing echo in speech. This approach helps create a dynamic model of echoed speech, as indicated by their findings.

3.1 Background

First, we'll go over some of the terminology and concepts associated with reverberant-noisy audio signal models. In a small room, the audio of a faraway speaker is an aggregation of echoes that have been damped and postponed by the room's many surfaces. Reverberation is the term used to describe this sonic alteration. In our investigation, we focus on a single room, hence the term "room impulse response" to describe the auditory reaction (RIR).

In mathematical terms, we use symbols s[t] for clean speech, y[t] for echo-filled speech, and h[t] for the room's characteristics. The echo-filled speech y[t] can be represented as follows:
$y[t] = h[t] * s[t]$

Where the symbol "*" in this context means a mathematical operation called convolution.

Many conventional methods for dereverberation and denoising, including numerous deep learning methods, rely on estimating a gain function h[t] for each frame and frequency bin. This gain function is then used to compute an estimate of the clean speech.

Here, we break down the gain function h[t] into two parts: one is the impulse response function h_d[t] for the direct sound, and the other is h_r[t] for the reverberation.

As a result, the reverberant sound is depicted by

$$y[t] = h_d[t] * s[t] + h_r[t] * s[t] = x[t] + r[t]$$

In our research, we aim to isolate the clean, echo-free information (the direct sound) x[t] from the recorded sound y[t] that includes echoes. It's important to note that x[t] can be quite different from the original clean speech s[t] because it undergoes changes over time and loses some energy during transmission. So, we use symbols x[t], y[t], and n[t] to refer to the direct sound, the echo-filled sound, and any background white noise, respectively. The combined sound with echoes and noise, which we call "z[t]," can be represented as follows:

$$z[t] = y[t] + n[t] = x[t] + r[t] + n[t]$$

We want to restore direct sound in order to improve reverberant-noisy communication x[t] based on its reverberant-noisy assessment z[t].

3.2 Proposed Methodology:

We take a time-domain audio signal recorded at 8 kHz and divide it into short frames. Each frame uses a 32-millisecond Hamming window and shifts every 8 milliseconds. Then, we apply a 512-point fast Fourier transform (FFT) to each frame, resulting in 257 frequency bands.

To simplify and capture the dynamic range, we use the cubic-root compressed magnitude spectrum of the echoed speech as a feature. These compressed magnitude spectral features at each time frame are represented as Y(m), which is a 257-dimensional vector.

Our algorithm takes the following sequential feature vectors as input:

$$Y = \{Y[1], Y[2], \dots Y[N]\}$$

Where "n" represents the total number of frames within a single statement..

The training objective is to obtain the cubic-root compressed amplitude spectrum of clean, echo-free audio. At frame m, we denote X[m] as the compressed amplitude range of this clean audio. The training goal is defined as follows:

$$X = \{X[1], X[2], \dots X[N]\}$$

The dereverberation problem is now expressed as a mapping problem from sequence to sequence, i.e.

$$\{Y[i]\} \rightarrow \{X[i]\}, \qquad i = 1,2,3, \dots, N$$

Deep learning methods leverage neural network (NN) models that have been trained on a dataset to enhance audio quality. In essence, these models serve as a transformation from an input feature vector xi to an output vector.

$$\mathcal{Y}_i = \mathcal{F}(\, x_i\,,\, \hbar_i\,, \Theta),$$

On the basis of the non-linear compound function $\mathcal{F}(.)$ The network structure and trainable factors determine Θ. The latent network states as an extra input $\hbar_i$ in recurrent neural networks, such as LSTMs, the previous frame is used to represent time background.

In the framework of methods built on deep learning, $\mathcal{G}_i(\hbar)$ from (3) is frequently referred to as a T-F filter that separates clear voice from noise. To predict these masks, the NN model can be taught in a controlled manner by minimising the mask approximation (MA) loss function.

$$\mathrm{M}_a = \frac{1}{\hbar}\sum \left(\mathcal{G}_i(\hbar) - \mathcal{G}_i^*(\hbar)\right)^2$$

$\mathcal{G}_i(\hbar)$: optimum mask values for the training targets

$\mathcal{G}_i^*(\hbar)$
: values for the estimated mask at the network output

The increased magnitude spectrum ought to be seamless due to similarities between consecutive frames. To make things smoother and more effective, we introduce a special type of layer called a 1-D convolutional layer with a small filter size. This layer is like a filter that helps our model understand and improve the central part of our data by looking at the surrounding context. It's like zooming in on important details to make our model work better. We use something called a "Rectified Linear Unit" or

ReLU for the final output because our data's values are all positive. ReLU helps us process this positive data effectively. To surmount the disadvantage of the conventional sequence network, three gates are introduced to the network's cell to enable the concept of memory. Whenever the cell reads data at each interval, a memory is maintained and refreshed.

LSTMs with four gate: forget (f), input (i), memory (c) and output gate (o)

Given an ancient remembrance C_{t-1}, the new cell memory C_t is calculated as:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Forget Gate: determines which information should be removed from the present memory

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

Memory Gate: creates fresh potential memory.

$$\tilde{C}_t = \tanh\,(W_c x_t + U_c h_{t-1} + b_c)$$

Input Gate: This gate decides how much of the information from the current prospective memory should be added to the new memory.

$$i_t = \sigma\,(W_i x_t + U_i h_{t-1} + b_i)$$

Output Gate: Specifies the amount of cell information is taken.

$$o_t = \sigma\,(W_o x_t + U_o h_{t-1} + b_o)$$

Figure 1 shows an architecture of proposed methodology.

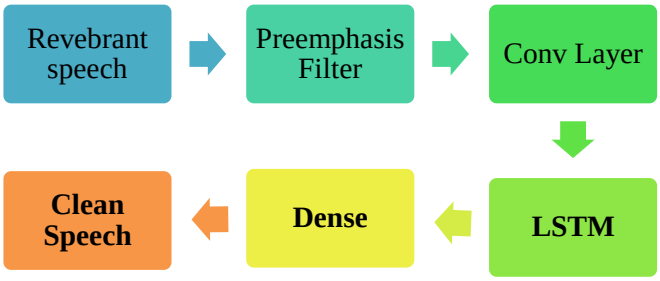

Fig 1: Architecture of the proposed methodology

The loss function we're using is called Mean Square Error (MSE).

$$\mathcal{L}[\mathrm{Y};\Theta] = \|\mathrm{X} - \mathcal{F}[\mathrm{Y}]\|_2^2$$

Where $\|.\|_2$ denotes the ℓ_2 norm and Θ denotes the parameters in the model $\mathcal{F}$, which are learned during training.

Algorithm for generating the cleaned speech

Input: Training dataset collect from various

sources $Train\ data\{\}$,$CNN\ Activation\ Function\{\}$, $Input\ Threshold, Th$

Output: Feature extraction from the convolutional and pooling layers

Step 1: Set input block of $Train\ data\{\}$, selective

$Activation\ Function\{\}$, $Epochs$,

Step 2 :

$Feature\ \{\} \leftarrow Extract\ Features\ Using\ Conv\ \&\ Pooling\ Layer$

$\{Train\ data\}$

Step 3 :

$Feature\ \{\} \leftarrow Optimised\ Features\ Using\ Dense\ layer\ \{Train\ data$

Step 4 :
$Training\ till\ Convergence\ reach \leftarrow LSTM\ layer\ \{\}$

Testing:

Input: $Test\ data\{\}$, $Input\ Threshold, Th$
$Trained\ Model$

Output: $Result\ Predicted\ Class$

Step 1: Read all t$Test\ data\{\}$, using the following function to verify the training rules,

$test\_Feature(data)$

$$= \sum_{m=1}^{n} (.\ Attribute\_Set[A[m] \dots\dots A[n] \leftarrow Test\_Data)$$

Step 2: Decide which features to use from the extracted attribute set $Extract\ Features\ Using\ Conv\ \&\ Pooling\ Layer$ and use the function below to create a feature map. Test_Feature_Map [t…..…n] = $\sum_{x=1}^{n}(t) \leftarrow test\_Feature(x)$

Test_FeatureMap [x] are the features that were chosen for the pooling layer.

Input features are extracted by the convolutional layer, which then sends those features along to the pooling layer are stored in Test_Feature_Map

Step 3: Now, in the sense layer, read the entire training dataset to construct the hidden layer for categorization of the entire test data.

$train\_Feature(data)$

$$= \sum_{m=1}^{n} (.\ Attribute\_Set[A[m] \dots\dots A[n] \leftarrow Train\_Data)$$

Step 4: Create the training map utilising the function described below using the input dataset.

Train_FeatureMap [t…..…n] = $\sum_{x=1}^{n}(t) \leftarrow train\_Feature(x)$

Train_FeatureMap[t] is the map of the hidden layer that produces the feature vector needed to construct the hidden layer. That uses train data to evaluate all test cases.

Step 5: We make a feature map to identify important parts of the data. Then, we assess how similar these important parts are to each other in the dense layer, particularly those chosen in the pooling layer.

Gen_weight

$$= CalcWeight\ (Test\_FeatureMap\ ||\ \sum_{i=1}^{n} Train\_FeatureMap[t])$$

Step 6: Check if the current weight is above or below the desired threshold to see if it's where you want it to be

$if(Gen\_weight >= qTh)$

Step 7: Out_List. add (trainF. class, weight)

Step 8: Go to step 1 and continue when Test_Data == null

Step 9 : Return Out_List

4. Experiment and Results

The proposed algorithm's performance is assessed using two objective measures frequently employed by the speech enhancement community, Specifically, log-spectral distance (LSD) and perceptual evaluation of speech quality (PESQ).

The scenarios listed below were considered:

i. Simulated reverberant signals with spatially white Gaussian noise;

ii. Simulated reverberant signals with air-condition noise; and

iii. Simulated reverberant signals with varying source-array Distance.

iv. Simulated reverberant signals with variation in reverberation time.

4.1 Datasets and Experimental Procedures

The proposed technique's performance is evaluated using the IEEE database (male and female speakers) [24]. A room with the dimensions 6.1x5.3x2.7 is simulated in order to produce RIRs. The array of four microphones is used, and the distance among two microphones is [3 4 3] cm . The RIR is produced by fixing the positions of the source and receiver microphones at 1 to 4 m apart. In this experiment, three incremental RT60 values of 0.3, 0.4, and 0.5 seconds are investigated. An image method [25] is used to produce the RIR.

Regarding the training and assessment data sets, reverberant and noisy signals are generated by mixing noise with the necessary SNR. Noise levels of 10 dB, 20 dB, and 30 dB SNR are used. In this paper, the sounds are captured at 8 kHz.

4,2 Evaluation Metrics

In this paper, by using a metrics called PESQ to evaluate how good the speech sounds when people talk. PESQ scores range from -0.5 to 4.5. To assess speech difference, the log-spectral distance is used. Because the aim is to eliminate room reverberation and background noise, using clear and echo-free speech as a reference signal to measure specific quantitative metrics. The results for PESQ and LSD are shown in the following tables. Reinforcement learning-based methods outperform the conventional Strategy in terms of PESQ and LSD on average, as well as for each specific SNR circumstance.

Table 1: PESQ of reverberant audio with spatially-white noise for a 3 m source-to-array distance and variable SNR

White Noise	10db	20db	30db
Unprocessed	1.5	1.88	2.01
DNN-WPE [6]	2.08	2.51	2.63
Proposed	**2.40**	**2.63**	**2.80**

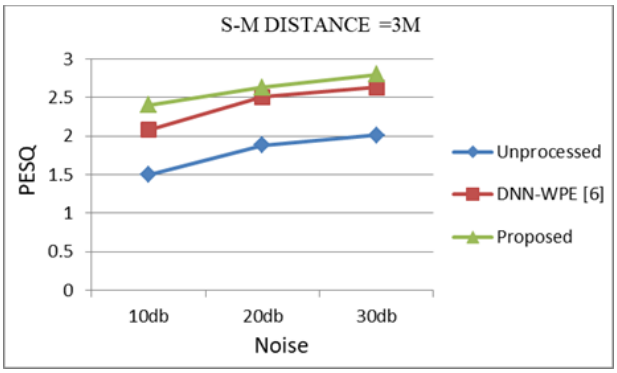

Fig 2: Plot of PESQ with white noise

Table 2: LSD of reverberant audio with spatially-white noise for a 3 m source-to-array distance and variable SNR

White Noise	10db	20db	30db
Unprocessed	16.34	9.4	5.23
DNN-WPE [6]	8.64	4.50	3.51
Proposed	**8.20**	**4.07**	**3.32**

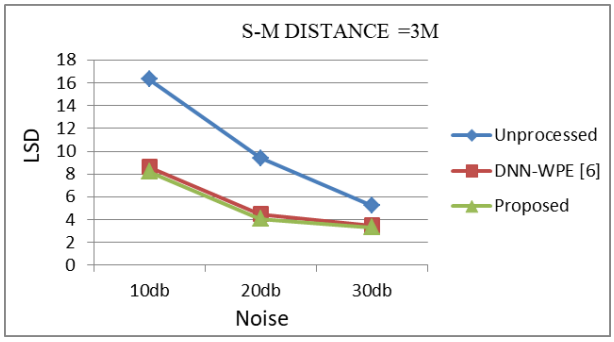

Fig 3: Plot of LSD with white noise

We experimented with four acoustics cases. The resulting PESQ scores and LSD are described in tables 1 to 8. The most important findings are emphasized in bold. The findings indicate that processed signals sound better (higher PESQ value) and have less sound difference (lower LSD) compared to raw signals in all the tables. The results shown in table 1, 2 and figure 2, 3 are for PESQ and LSD values with different SNR with spatially-white noise and a speaker-array separation of 3 m. It demonstrates improvement in speech quality with increase in PESQ value and reduction in LSD value in comparison with unprocessed signal and reference technique.

Table 3 :PESQ of reverberant audio air-conditioning noise for a source to array distance of 3 m and varying SNR

AC Noise	10db	20db	30db
Unprocessed	1.75	2.1	2.22
DNN-WPE [6]	2.24	2.53	2.69
Proposed	**2.48**	**2.90**	**2.96**

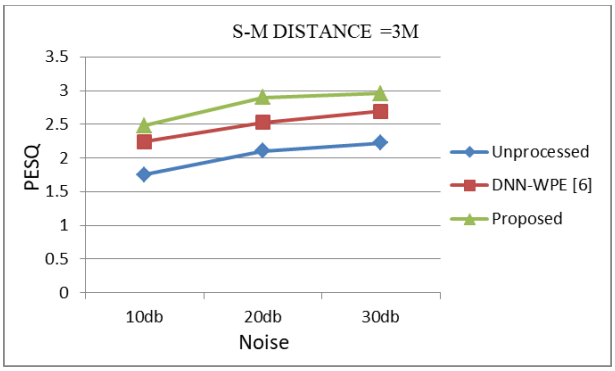

Fig 4: Plot of PESQ with air-conditioning noise

Table 4 :LSD of reverberant audio air-conditioning noise for a source to array distance of 3 m and varying SNR

AC Noise	10db	20db	30db
Unprocessed	9.52	6.31	4.78
DNN-WPE [6]	6.76	4.90	3.90
Proposed	5.39	3.84	3.25

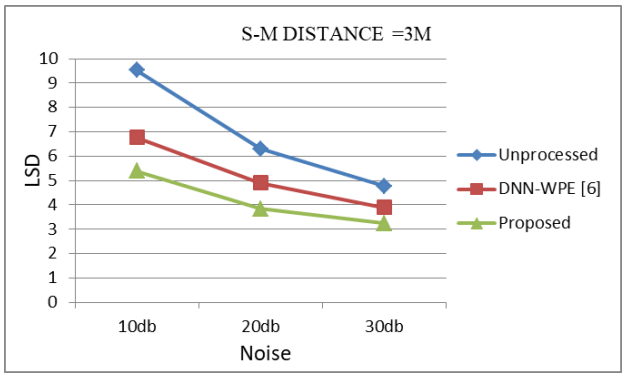

Fig 5: Plot of LSD with air-conditioning noise

Table 3, 4 and figure 4, 5 shows the results with air-conditioning noise for different SNR values illustrating better speech quality with increase in PESQ and reduction in LSD values compared with references values.

Table 5 : PESQ of reverberant audio signal for varying source to array distance

S-M Distance	1m	2m	3m	4m
Unprocessed	2.74	2.35	2.26	2.27
DNN-WPE [6]	3.11	2.76	2.71	2.68
Proposed	3.70	3.11	2.99	2.98

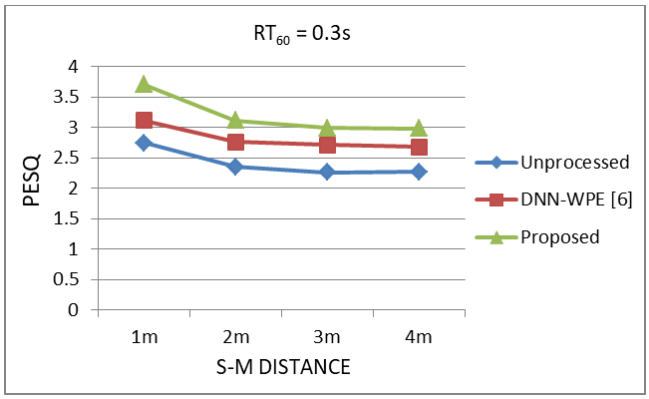

Fig 6: Plot of PESQ for S-M distance of 1m to 4m

Table 6 : LSD of reverberant audio signal for varying source to array distance

S-M Distance	1m	2m	3m	4m
Unprocessed	2.96	3.89	4.29	4.47
DNN-WPE [6]	2.43	3.37	3.71	4.06
Proposed	2.68	3.05	3.27	3.17

Table 5, 6 and figure 6, 7 depicts the results for variations in source to microphone array distance. The spacing between the source and the microphone is varied from 1m and 4m. In table 5 and figure 6 PESQ of proposed technique is increased in comparison with unprocessed signal and reference technique for distance 1m to 4m. As depicted in table 6 and figure 7, for distance 1m to 4m LSD is decreased in comparison with reference methods.

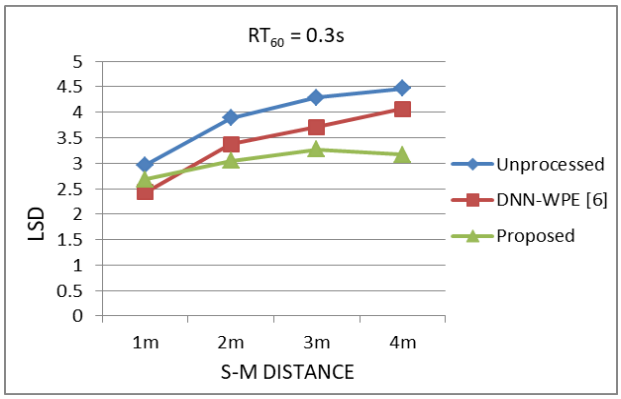

Fig 7: Plot of LSD for S-M distance of 1m to 4m

Table 7:PESQ of reverberant audio signal for varying reveberation time

RT60	0.3s	0.4s	0.5s
Unprocessed	2.74	2.35	2.26
DNN-WPE [6]	3.16	2.80	2.59
Proposed	3.70	3.12	3.00

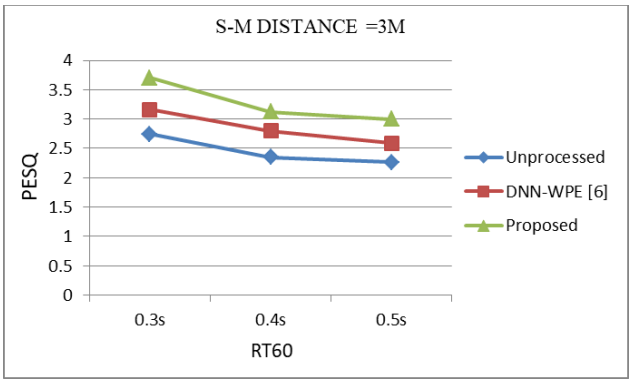

Table 8:LSD of reverberant audio signal for varying reveberation time

RT60	0.3s	0.4s	0.5s
Unprocessed	2.96	3.89	4.29
DNN-WPE [6]	2.82	3.31	3.76
Proposed	2.38	3.00	3.27

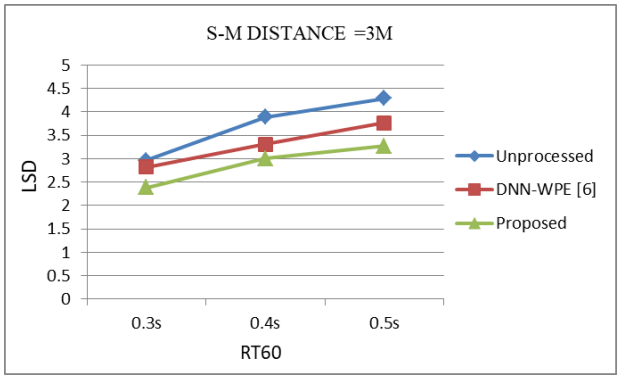

Fig 9: Plot of LSD with different values of RT60

In table 7, 8 and figure 8, 9 the variation in reverberation time is considered. The RT60 duration used in this is 0.3, 0.4, and 0.5 seconds. The PESQ value of proposed algorithm is increased in comparison with unprocessed signal and reference technique as depicted in table 7 and figure 8. The LSD value of proposed algorithm is decreased in comparison with unprocessed signal and reference technique as depicted in table 8 and figure 9. So for all values of RT60, PESQ is increased and LSD is decreased in comparison with reference methods indicating increase in speech quality.

5. Conclusion

Robust speech processing tasks frequently necessitate speech dereverberation and denoising. The state-of-the-art result in multi-channel speech dereverberation is provided by supervised deep learning (DL) models. The suggested method uses an approximation of a complicated time-frequency filter to help recover the original formant structure. This article proposes a Reinforcement learning-

based technique for simultaneous training of voice denoising and dereverberation. In comparison to other deep learning-based generative adversarial frameworks, the suggested system consistently outperforms the competition across a varied range of noise levels, reverberation time and source to microphone array distance. When compared with the DNN-WPE [6] algorithm, the PESQ values for white noise are increased by 15.38% and the LSD values reduce by 5.09% with the proposed algorithm. In the same way, the PESQ values for the air conditioner noise are increased by 10.71% and the LSD values decreases by 20.00% at 10 db SNR The result also demonstrates that the proposed algorithm outperforms the DNN-WPE [6] algorithm, even with increases in source to microphone distance and reverberation time. The experimental results indicate that the proposed algorithm works better than the existing algorithm..

Author contributions

Seema Arote: Conceptualization, Methodology, Software, Field study **Vijay Mane:** Data curation, Writing-Original draft preparation, Software, Validation, **Dattaray Bormane:** Field study, Visualization **Shakil Shaikh:**, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] S. Gul et.al, “Preserving the beamforming effect for spatial cue-based pseudo-binaural dereverberation of a single source”, Computer Speech & Language, Volume 77, 2023, 101445, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2022.101445.

[2] J. Bruyninckx, “Tuning the office sound masking and the architectonics of office work”, Received 06 Dec 2021, Accepted 22 Dec 2022, Published online: 03 Feb 2023 https://doi.org/10.1080/20551940.2022.2162765

[3] Y. Li, et.al, "A Composite T60 Regression and Classification Approach for Speech Dereverberation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 1013-1023, 2023, https://doi.org/10.1109/TASLP.2023.3245423

[4] N. Prodi et.al, “Comparing the effects of scattered and specular sound reflections on speech intelligibility in rooms, Building and Environment”, Volume 228, 2023, 109881, ISSN 0360-1323, https://doi.org/10.1016/j.buildenv.2022.109881

[5] L. Passos et.al, “Multimodal audio-visual information fusion using canonical-correlated Graph Neural Network for energy-efficient speech enhancement”, Information Fusion, Volume 90, 2023, Pages 1-11, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2022.09.006 .

[6] J. Sheeja et.al, “Speech dereverberation and source separation using DNN-WPE and LWPR-PCA”. Neural Comput & Applic (2023). https://doi.org/10.1007/s00521-022-07884-0

[7] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Processing Letters, vol. 21, no. 1, pp. 65–68, 2013.

[8] D. Wang and J. Chen, “Supervised speech separation based on deep learning: an overview,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, 2018.

[9] Z. Jin and D. Wang, “A supervised learning approach to monaural segregation of reverberant speech,” IEEE Trans. Audio Speech Lang. Process., vol. 17, no. 4, pp. 625–638, 2009.

[10] X. Li, J. Li, and Y. Yan, “Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions,” in Proceedings of the INTERSPEECH, pp. 1203–1207, Stockholm, Sweden, August 2017.

[11] M. Kolbaek, D. Yu, Z.-H. Tan et al., “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901–1913, 2017.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, “Phase sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 708–712, Brisbane, Australia, April 2015.

[13] D. S. Williamson and D. L. Wang, “Speech dereverberation and denoising using complex ratio masks,” in Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5590–5594, New Orleans, LA, USA, March 2017

[14] Z. -Q. Wang and D. Wang, "Deep Learning Based Target Cancellation for Speech Dereverberation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 941-950, 2020, https://doi.org/10.1109/TASLP.2020.2975902 .

[15] Hangting Chen, Pengyuan Zhang,"A dual-stream deep attractor network with multi-domain learning for speech dereverberation and separation", Neural Networks, Volume 141,2021,Pages 238-248, ISSN 0893-6080

[16] H Wanget.al, "TeCANet: Temporal-Contextual Attention Network for Environment-Aware Speech Dereverberation", Audio and Speech Processing,arXiv:2103.16849

[17] Y. Fu et.al.,"DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation," 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 857-864, https://doi.org/10.1109/SLT48900.2021.9383604 .

[18] C. Fan et.al," Simultaneous Denoising and Dereverberation Using Deep Embedding Features,Audio and Speech Processing,arXiv:2004.02420,6 Apr 2020

[19] H. Li et.al, "Robust Speech Dereverberation Based on WPE and Deep Learning," 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020, pp. 52-56.

[20] Xiao et al. Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation. EURASIP J. Adv. Signal Process. 2016, 4 (2016). https://doi.org/10.1186/s13634-015-0300-4

[21] K. Han, et.al, "Learning Spectral Mapping for Speech Dereverberation and Denoising," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 6, pp. 982-992, June 2015, https://doi.org/10.1109/TASLP.2015.2416653 .

[22] Y. Masuyama, et.al., "End-to-End Integration of Speech Recognition, Dereverberation, Beamforming, and Self-Supervised Learning Representation," 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023, pp. 260-265, https://doi.org/10.1109/SLT54892.2023.10023199 .

[23] S. Gul et.al.,Preserving the beamforming effect for spatial cue-based pseudo-binaural dereverberation of a single source, Computer Speech & Language, Volume 77, 2023, 101445, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2022.101445 .

[24] "IEEE Recommended Practice for Speech Quality Measurements," in IEEE Transactions on Audio and Electroacoustics, vol. 17, no. 3, pp. 225-246, September 1969

[25] Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer. 4(65), 943–950 (1979).